Common Scab Detection on Potatoes Using an Infrared Hyperspectral Imaging System

Angel Dacal-Nieto[1], Arno Formella[1], Pilar Carrión[1],
Esteban Vazquez-Fernandez[2], and Manuel Fernández-Delgado[3]

[1] Computer Science Department, Universidade de Vigo,
Campus As Lagoas 32004 Ourense, Spain
angeldacal@uvigo.es
[2] GRADIANT, Galician R&D Center in Advanced Telecommunications, Spain
[3] Centro de Investigación en Tecnoloxías da Información (CITIUS),
Universidade de Santiago de Compostela, Spain

Abstract. The *common scab* is a skin disease of the potato tubers that decreases the quality of the product and influences significantly the price. We present an objective and non-destructive method to detect the common scab on potato tubers using an experimental hyperspectral imaging system. A supervised pattern recognition experiment has been performed in order to select the best subset of bands and classification algorithm for the problem. Support Vector Machines (SVM) and Random Forest classifiers have been used. We map the amount of common scab in a potato tuber by classifying each pixel in its hyperspectral cube. The result is the percentage of the surface affected by common scab. Our system achieves a 97.1% of accuracy with the SVM classifier.

Keywords: Hyperspectral, Infrared, Potato, SVM, Random Forest.

1 Introduction

Detecting and identifying defects and diseases in potato tubers (*Solanum tuberosum*) continue to be an important challenge for food engineering and automation. Industry uses a large variety of technologies and computer vision methods have been a specially successful choice. Nevertheless, some new technologies should be taken into account for improving non-destructive potato quality assessment.

The importance of the potato industry is extreme, since potatoes are still one of the most consumed products in the world; they are the world's fourth largest food crop. The annual production is 325 million tons and it moves an amount of global transactions of about 6 billion US dollars (2007 data). Thus, the world potato average consumption is 31 kg per capita and year [1].

Hyperspectral imaging is an emerging technology originally designed for military remote satellite inspection [2], but also used for remote sensing, astronomy and earth observation. It is also a reliable approach to classical spectroscopy, because despite a little loss of accuracy, an object can be analysed in significantly less time, in a non-destructive way.

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part II, LNCS 6979, pp. 303–312, 2011.

The scientific community has started to show its interest in the last years in hyperspectral imaging possibilities for food quality [3]. Regarding the research in potato quality assessment, there are systems to predict the water content in potatoes using classical spectroscopy techniques [4]. Some other contributions are oriented to the detection of clods between a set of potato tubers using hyperspectral imaging [5]. Finally, there are contributions [6] that investigate composition characteristics from potato tubers like water, starch and proteins, using invasive spectroscopy techniques, meanwhile others [7] use NIR spectroscopy to predict specific gravity and dry matter in potatoes. Using other optical spectral methods [8], there are contributions for the detection of common scab, dry rot, gangrene, and other diseases, using wavelength ranges between 590 nm and 2030 nm, and getting accuracies up to 83%. Unfortunately, these systems are either destructive or they can not be easily included in classical machine vision developments in order to use the same image acquisition for all the processes.

Our objective is to map the common scab affected areas in potato tubers. This has been achieved in the past by using different technologies, as it has been described before. However, mapping the common scab is not only a required objective itself: we also use this mapping as a preprocessing stage into a wider potato inspection system, which detects internal and external defects and diseases. Some of these diseases require a morphological study, so hyperspectral technology seems to be the best approach. It would be interesting to provide a solution using the same image acquisition system, in order to unify the inspection process, so hyperspectral imaging has been also the selected technology to solve the common scab mapping problem.

Our solution is objective, automatic and non–destructive. Nevertheless, this choice makes difficult the comparison with other common scab detection methods. In fact, there are not hyperspectral solutions yet for detecting common scab, due to the novelty of the technology. Moreover, previous spectral contributions used different wavelength ranges, or performed a combined searching of other diseases, so a partial comparison is presented.

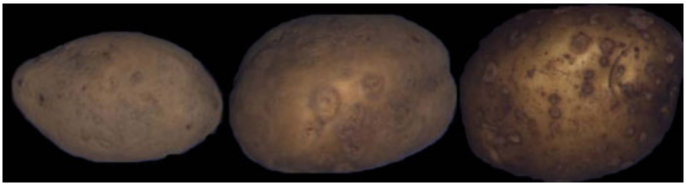

Fig. 1. Three examples of common scab affected potatoes

2 Image Acquisition System

The concept of hyperspectral imaging is to perform a spectroscopic analysis of the light reflected or transmitted by the object of interest. We couple a spectrograph and a matrix camera to obtain both spectral and spatial information.

The camera is a Xenics Xeva 1.7–320 with an InGaAs 320 × 256 pixel sensor and USB connection ("http://xenics.com"); the spectrograph is a Specim Imspector N17E ("http://specim.fi"). Both are sensitive from 900 nm to 1700 nm. The system has also three 50 W AC halogen lamps placed in the inspection plate. The illumination is diffused by the reflection in a plastic dome over the plate.

The spectrograph has a linear input (one pixel height), where the x-axis represents the same x-axis (spatial) of the object. The y-axis (spectral) is then *studied* to obtain how every pixel in the row varies along the spectral range.

With one spectral image, we are inspecting only one spatial line, so we need to perform the inspection over the whole object. This is accomplished by joining a rotatory Mirror Scanner ("http://specim.fi") to the spectrograph. It is based on performing the mirror rotation over the object, taking care of synchronization between mirror stepping and image acquisition (Figure 2). Finally, the obtained images are transposed in order to obtain the hyperspectral cube (Figure 3).

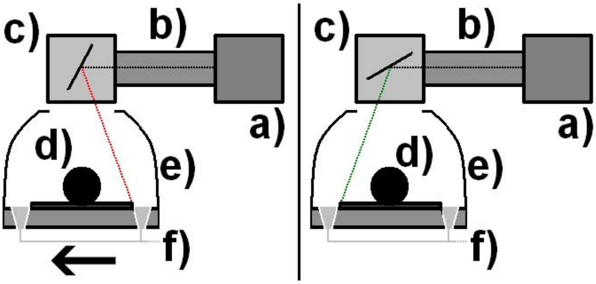

Fig. 2. Left: scanning initial position at 70°. Right: scanning final position at 110°. The arrow shows the direction of scanning. Hyperspectral system scheme: a) camera, b) spectrograph, c) mirror scanner, d) object, e) diffuse chamber, f) halogen lamp.

To sum up, our system obtains 320 spectral images (320 × 240 pixels), that are transposed into hyperspectral cubes formed by 256 images with 320 × 320 pixels, corresponding to 256 consecutive wavelengths from 900 nm to 1700 nm.

3 Experiment

We use a set of 234 potato tubers (variety Agria) from Xinzo de Limia (Spain), with different degrees of common scab incidence, that have been collected from some potato packing companies during the 2009 harvest.

3.1 Segmentation

In every hyperspectral cube, we need to segment the potatoes from the background for later mapping tasks. We segment only one image from the hyperspectral cube (the wavelength 980 nm has been found after several performance tests). We obtain a mask that is applied in the rest of the cube.

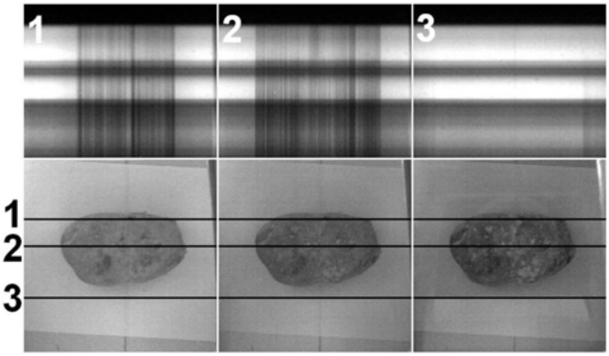

Fig. 3. Up: Three spectral images taken from different lines of the object. Down: 978 nm, 1173 nm, and 1608 nm spatial images.

Segmentation runs in several steps (Figure 4), helped by the open source library OpenCV [9]. First, we binarize the image using Otsu's method [10] that calculates the optimum binarization threshold using a probabilistic analysis of the image. Then, a Gaussian blurring clusters the noise in the image. Another binarization is needed before a connected-component labelling, performed to remark contiguous areas in the image. At this point, we know that the blob with the largest area (excluding the background) is the potato. We select this blob and create the mask used to segment all the images in the hyperspectral cube.

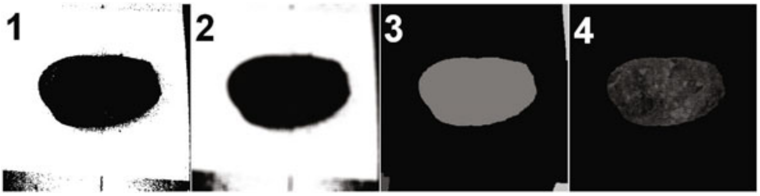

Fig. 4. 1: Binarization using Otsu's method. 2: Smooth operation. 3: Blob analysis. 4: Example image after applying the full mask.

3.2 Feature Extraction

In our problem, we have to distinguish two classes: *common scab* and *healthy*. To create a dataset with both common scab affected samples and healthy samples, experts helped us to identify which portions (ROI) were affected and which were not. Some hyperspectral cubes provided more samples (especially those more affected by common scab) meanwhile others provided just one healthy sample.

Note that every hyperspectral cube consists of 256 images that correspond to the 256 bands of the hyperspectral system. For this reason, when we select a ROI, we are not selecting just a rectangle of pixels, but that rectangle all over the 256 images that are part of the hyperspectral cube.

When selecting the ROI, the average intensity value of the pixels in the ROI is calculated for each band. Hence, every sample (independently of its size) is represented with 256 attributes and an extra attribute that denotes the class (common scab or healthy). Eventually, we have obtained 649 samples (208 corresponding to common scab class and 441 corresponding to healthy class).

The samples can be visualized in a chart where the x-axis represents the wavelength range and the y-axis the grey level in the band, which is actually the arithmetic mean of pixels in every band of the ROI. In the Figure 5 we can see ROI selection of two samples and the corresponding luminance charts.

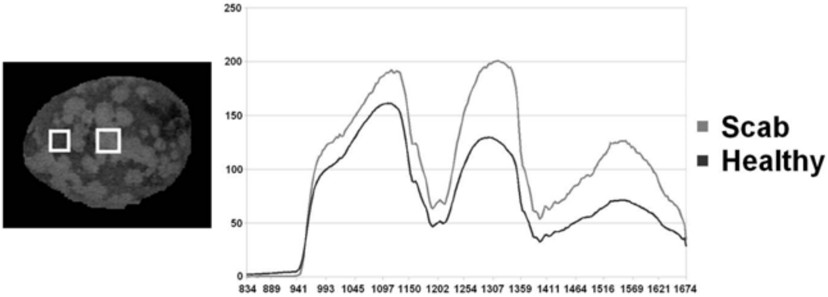

Fig. 5. Left: healthy and common scab affected (the brightest) ROI's. Right: Luminance charts from two different samples. The x-axis represents the wavelength. The y-axis represents the average grey level in the ROI, for each band.

3.3 Feature Selection

Feature selection is a common task in pattern recognition, specially if the initial number of features is high. With less features, the learning process is faster and the generalization capabilities of the classifier are improved. In our case feature selection is a fundamental step to decrease the overall execution time, identifying which wavelengths are sufficient to solve the common scab detection problem.

We have tested some techniques regarding spectral bands selection on hyperspectral imaging systems, implemented on Weka [11]: Genetic Search [12] (which selects 11 bands), Scattered Search [13] (11 bands), Greedy Stepwise [14] (5 bands), Linear Forward Selection (LFS) [15] (7 bands), and Correlation-based Feature Subset Selection (CFS) [16], (6 bands). Note that with CFS, three contiguous zones have been selected: 1300 nm–1303 nm, 1336 nm–1342 nm and 1503 nm. Techniques such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are the most commonly used unmixing techniques in spectral imaging. However, these algorithms do not reduce the number of wavelengths needed, rather they generate a linear combination of the 256 features into a new feature space. This is the reason why only feature selection operations are interesting in this research.

To summarize, after this step of the experiment, we provide six datasets to the classification procedure: *genetic*, *scattered*, *greedy*, *LFS*, *CFS* and *full*.

3.4 Classification Algorithms

We present results for two classification algorithms: Random Forest (RF) and Support Vector Machines (SVM). Other classifiers have been tested in a preliminary stage (Logistic Regression, MLP and k-NN), but they performed poorly.

A Random Forest [17] is a collection of trees that classifies individually an input sample, and then evaluates the individual responses of the trees to output the mostly voted class. We used the OpenCV implementation of RF, tuning the m_{try} parameter, which is the number of features to be used in random selection.

SVM find the optimal hyperplane over a high dimensional space where the feature vectors have been mapped using a kernel function (Gaussian in our case). We can tune its behaviour with the regularization parameter (also known as cost, or C), which is not very relevant for the results [18], and the kernel spread (γ), with high relevance on the classification accuracy. SVM has been introduced in our system with the library LibSVM [19].

3.5 Classification Evaluation Procedure

For each dataset, we evaluated the classification algorithms using a method based on randomly generating 10 permutations of the dataset, so that each permutation has the same samples, but differently ordered. Then, each permutation is divided into three parts: *training* (50% of the samples), *validation* and parameter tuning (25% of the samples), and *test* (remaining 25%). The samples are normalized (zero mean and standard deviation one) to avoid that attributes in greater numeric ranges influence excessively over those with smaller variation.

For each combination of tunable parameters and for each permutation, we train a classifier using the training sets. Then, we test its performance by using the validation sets. We selected the parameter values which provide the best average accuracy over the 10 permutations.

In the case of RF, the default value of m_{try} is $\sqrt{p}$, being p the number of features of the problem. We follow a parameter tuning as being suggested by [20]. We use different values of m_{try}: $m_{\text{try}} = p^0$, $m_{\text{try}} = \sqrt{p}$, $m_{\text{try}} = p/4$ and $m_{\text{try}} = p/2$. The rest of the parameters have been established as [20] recommends. Thus, the number of trees has been set to 500, since it is enough, and there is no penalty for having an excessive number of trees.

In the case of SVM, we try pairs of (C, γ) using exponentially growing sequences for C and γ [19]. Thus, we use $C = 2^n, n = -5..14$ and $\gamma = 2^n, n = -15..3$, which gives 380 combinations. A finer adjustment has been discarded after some preliminary tests.

Finally, for each permutation, we train the classifier using the training sets tuned with the best parameters found, evaluating its accuracy on the test sets.

Note that using more permutations prevents unfair divisions of the dataset. For example, using only one permutation, if all the *easy-to-classify* samples are filled in the test set, it would cause unfairly good results. Additionally, each dataset has also been evaluated using leave-one-out cross-validation (loocv).

3.6 Affected Surface Measurement

Once being able to classify, the objectives are, for each potato (actually for each hyperspectral cube), obtaining an image that marks which zones are affected and computing the percentage of common scab affected surface.

First, we segment the potato, removing the background from the hyperspectral cube. Then, each pixel in the hyperspectral cube is classified individually (excluding the background, that has been localized previously in the segmentation step). Each hyperspectral pixel has (in our system) 256 values: however, depending on the feature selection procedure, we will have only a few of them. With the information of membership of each pixel to one class or another, we create a common scab map image. To reduce noise in the final map, a closing operation is performed followed by an opening operation with the same kernel. Finally, we calculate the percentage of the affected surface.

The objective of our system is to inspect 20 Kg samplings. Each potato will be inspected only by one side. Inspecting such an amount of potatoes averages individual errors, since we provide a statistical measurement. The result will be the average affected surface in the whole 20 Kg sampling.

4 Results and Discussion

The results of all datasets and classifiers can be seen in Table 1 including the leave-one-out cross-validation. Support Vector Machines show to be more effective than Random Forest in all the datasets. On the other hand, the CFS dataset seems to have the better subset of features, so that the pair SVM and CFS dataset is the best option to solve our problem.

Table 1. Accuracy (in %) for each dataset and classification algorithm

Classifier	Dataset	Accuracy %	loocv Acc. %	Best params.		Valid. Acc. %	Bands
					m_{try}		
RF	full	95.4	96.1		$\sqrt{p}$	95.6	256
	genetic	94.3	96.2		$p/2$	93.6	11
	scattered	93.8	96.9		$p/4$	95.0	11
	greedy	95.9	97.4		$p/2$	96.7	5
	LFS	94.5	96.8		$\sqrt{p}$	95.2	7
	CFS	95.8	96.5		$\sqrt{p}$	96.1	6
				C	γ		
SVM	full	96.2	96.6	2^{-2}	2^{-10}	96.4	256
	genetic	96.0	96.8	2^{5}	2^{-10}	96.5	11
	scattered	95.7	96.9	2^{5}	2^{-2}	96.5	11
	greedy	96.7	96.9	2^{12}	2^{-15}	97.0	5
	LFS	96.0	97.7	2^{7}	2^{-1}	96.9	7
	CFS	**97.1**	**98.0**	2^{11}	2^{-5}	**97.4**	6

As commented in Section 3.3, PCA and similar methods have been analysed but considered not adequate. However, some preliminary work has been done to check their performance. Thus, a new dataset has been created using Weka, after applying the PCA method to the *full* dataset. The loocv results show that this dataset gets a 95.2% of accuracy with RF, and approximately a 96% of accuracy using SVM. These results are 2 points under the feature selection algorithms results, and even worse the *full* dataset.

Now we are going to study further the best dataset–classifier pair. The best combination of parameters found was to be $C = 2^{11}$ and $\gamma = 2^{-5}$. The confusion matrix can be seen in Table 2. Note that these results were obtained using the test sets, composed by 25% of the samples (162 in our case). This is an average confusion matrix taking into account the ten permutations.

Table 2. Average confusion matrix obtained with the CFS dataset using SVM

Classified as *Real*	Common Scab	Healthy
Common Scab	48.3	3.1
Healthy	1.6	109

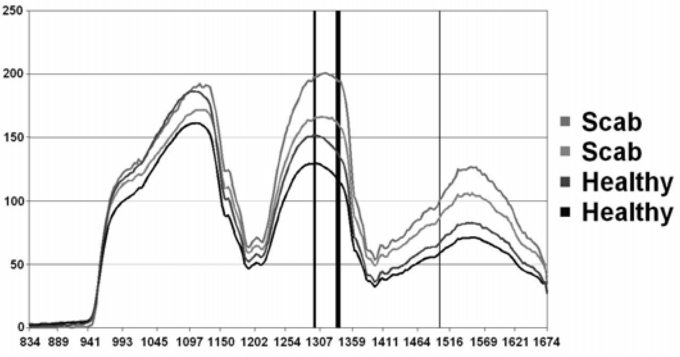

Fig. 6. Four samples, two from each class. Columns in grey mark the zones used by the CFS dataset. The rest of the bands were not selected. The x-axis represents the wavelength. The y-axis represents the grey level.

Four samples (two of each class) are presented in their luminance chart in Figure 6. Columns in black mark zones being selected in the CFS dataset. Previous contributions [21] show that at wavelengths greater that 1100 nm, where absorption by water dominates, the reflectance increases due to dehydration in the affected area, as in the case of common scab. Our automatically selected wavelengths lie in that range. However, by the moment it is impossible to compare our results with other common scab detection methods, since they use different wavelength ranges, or searched for other diseases in the same experiment.

5 Conclusions

Hyperspectral imaging has shown to be an good technology applied to food quality assessment. We have used an objective and non-destructive infrared hyperspectral system to identify the surface affected by common scab on potatoes.

Several feature selection algorithms have been tested, showing that this is a critical step to increase the system speed, because only 6 bands achieve the best accuracy. The selected bands with the CFS method (1300 nm, 1303 nm, 1336 nm, 1339 nm, 1342 nm and 1503 nm) provide enough information to classify common scab and healthy surface with a 97.1% of accuracy using the SVM classifier (tuned with $C = 2^{11}$ and $\gamma = 2^{-5}$).

This information could be useful for the designing of a specific multispectral image acquisition system, which would not have any mechanical device to move the camera or the object, because images could be captured within a reduced and specifically chosen group of wavelengths (in our case around 1301 nm, 1339 nm and 1503 nm). Hyperspectral cube reconstruction would not be needed any more. Hence, the time spent in an acquisition session would be considerably reduced.

The system will be used as a preprocessing step to remove the common scab for improving other disease identification algorithms on potatoes, as hollow heart, or the dry matter estimation amount. In future work, it would be interesting to evaluate the system with other potato varieties. On the other hand, methods like LDA should be tested in order to compare with the feature selection methods used in this paper. Finally, the relationship between the wavelengths selected with the best dataset and the biological components of common scab should be researched. This could be achieved by using a different image acquisition system (i.e. sensitive from 500 nm to 2000 nm), in order to compare our results with the obtained in [8].

Acknowledgements. This work was partially supported by the Spanish regional government of the “Xunta de Galicia” (projects PGIDIT08TIC004CT and PGIDIT08MMA010402PR). We want to acknowledge the “Laboratorio Oficial de Metroloxía de Galicia (LOMG)” and the “Centro Tecnolóxico da Carne (CTC)” for their helpful collaboration.

References

1. Potato World - International Year of the Potato (2008), `http://www.potato2008.org/en/world/index.html` (accessed January 01, 2011)
2. Goetz, A., Vane, G., Solomon, J.E., Rock, B.N.: Imaging spectrometry for earth remote sensing. Sci. 228(4704), 1147–1153 (1985)
3. Sun, D.: Hyperspectral Imaging for Food Quality Analysis and Control. Academic Press, Elsevier, San Diego, California (2009)
4. Singh, B.: Visible and near-infrared spectroscopic analysis of potatoes. M.Sc. Thesis. McGill University, Montreal, PQ, Canada (2005)
5. Al-Mallahi, A., Kataoka, T., Okamoto, H., Shibata, Y.: Detection of potato tubers using an ultraviolet imaging-based machine vision system. Biosyst. Eng. 105, 257–265 (2009)

6. Buning-Pfaue, H.: Analysis of water in food by near-infrared spectroscopy. Food Chem. 82, 107–115 (2003)
7. Kang, S., Lee, K., Son, J.: On-line internal quality evaluation system for the processing potatoes. In: Food Process. Autom. Conf. Proc., Providence, Rhode Island (2008)
8. Porteous, R.L., Muir, A.Y., Wastie, R.L.: The Identification of Diseases and Defects in Potato Tubers from Measurements of Optical Spectral Reflectance. J. Agric. Eng. Res. 26, 151–160 (1981)
9. Bradski, G., Kaehler, A.: Learning OpenCV: Computer Vision with the OpenCV Library. O'Reilly Media, Sebastopol (2008)
10. Otsu, N.: A threshold selection method for gray level histograms. IEEE Trans. Syst. Man Cybern. 9, 62–66 (1979)
11. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update. SIGKDD Explorations 11(1) (2009)
12. Goldberg, D.: Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, Reading (1989)
13. García-López, F., García-Torres, M., Melián-Batista, B., Moreno-Pérez, J.A., Moreno-Vega, J.M.: Solving feature subset selection problem by a Parallel Scatter Search. Eur. J. Oper. Res. 169(2), 477–489 (2008)
14. Weihs, C.: Multivariate exploratory data analysis and graphics, a tutorial. J. Chemom. 7, 305–340 (1993)
15. Guetlein, M., Frank, E., Hall, M., Karwath, A.: Large Scale Attribute Selection Using Wrappers. In: Proc. IEEE Symposium on Computational Intelligence and Data Mining, pp. 332–339 (2009)
16. Hall, M.: Correlation-based Feature Subset Selection for Machine Learning. Hamilton, New Zealand (1998)
17. Breiman, L.: Using Iterated Bagging to Debias Regressions. Mach. Learn. 45, 261–277 (2001)
18. Valentini, G., Dietterich, T.G.: Bias-variance analysis of support vector machines for the development of SVM-based ensemble methods. J. Mach. Learn. Res. 5, 725–775 (2004)
19. Chang, C.C., Lin, C.J.: LIBSVM:a library for support vector machines (2008), `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`
20. Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P., Feuston, B.P.: Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. J. Chem. Inf. Comput. Sci. 43, 1947–1958 (2003)
21. Gunasekaran, S., Paulsen, M.R., Shove, G.C.: Optical methods for nondestructive quality evaluation of agricultural and biological materials. J. Agr. Eng. Res. 32, 209–241 (1985)